Information Architecture for Interactive Archives (IAIA) Team

Chiu Wiegand

Justin Boblitt

CCMC

Introducing the Team

- Leads: Chiu Wiegand, Daniel Heynderickx, Darren De Zeeuw, Todd King
- International collaboration effort: SPASE experts, IMPEx experts, Virtual observatories and data centers across the globe
  - Full list of participants: https://ccmc.gsfc.nasa.gov/challenges/IAIAinfo/Participants.php
- Contact:
  - ccmc-iaia@googlegroups.com
  - SLACK: ccmc-collab.slack.com
- https://ccmc.gsfc.nasa.gov/assessment/topics/data.php

Mission Statement

- Facilitate the development of a global network of distributed web-based resource for the purpose of model-data comparison
- Focused Area:
  - Metadata standards/data model to describe observation and model metadata (implementation of SPASE with IMPEx extension)
  - Data discovery and access via standard Application Programming Interfaces (APIs)
  - Next generation interpolation libraries/approach
  - Web-based visualization tool

Metadata and Why?

- What?
  - Metadata is data that provides information about your data
- Why?
  - Data discovery, model-data comparison, validation of models
  - If you have generated or used any data sets for your projects, you know how important metadata is
    - Usually include it in the header or comment section of data files
  - Why do we need to use a standard for metadata?
    - Common language and format for ease of data comparison and sharing
    - Introducing SPASE with IMPEx extension

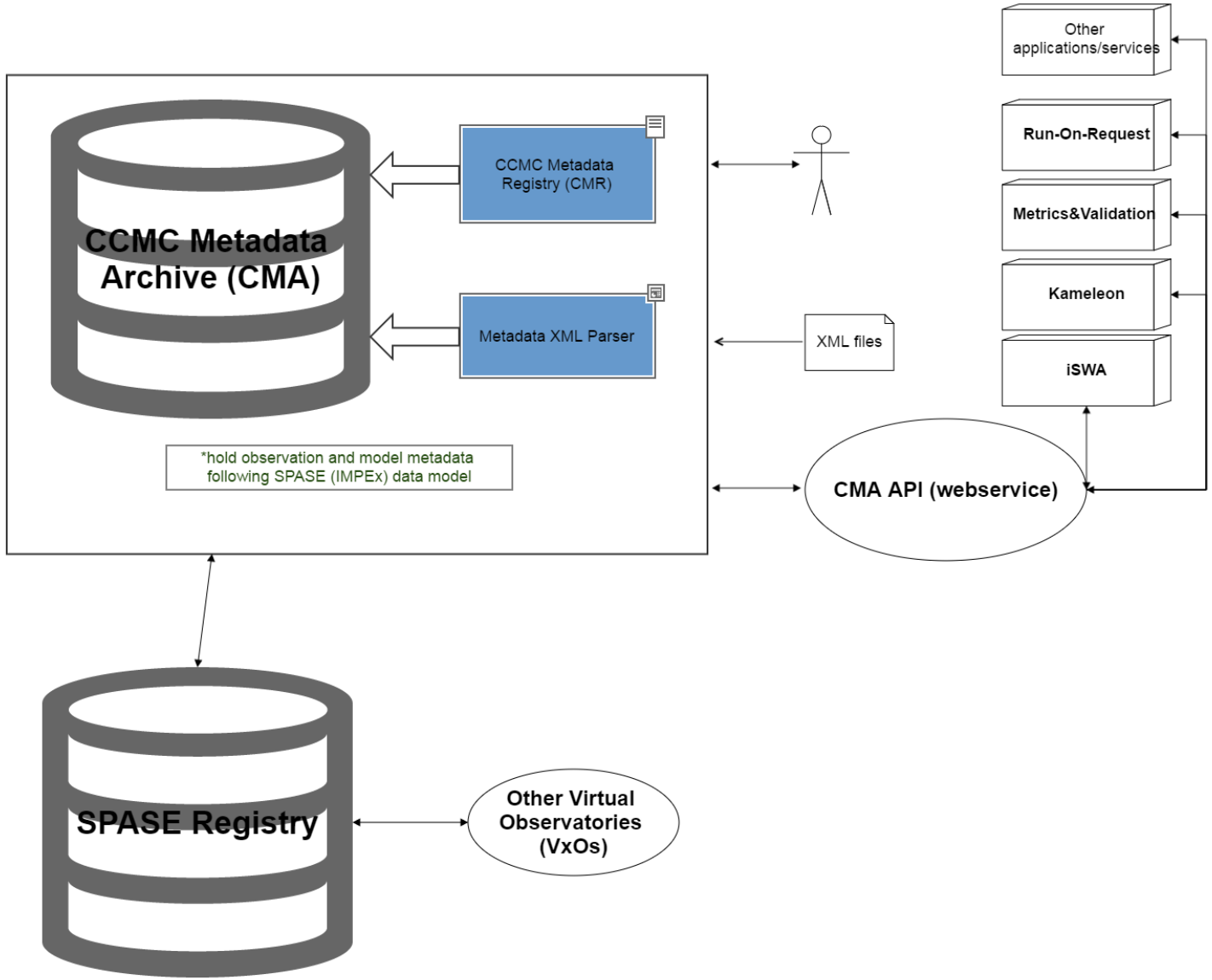
The Space Physics Archive Search and Extract (SPASE)

- The SPASE effort is a Heliophysics community-based project with the goals of:
  - **Facilitating data search and retrieval** across the Space and Solar Physics data environment with a common metadata language
  - Defining and maintaining a **standard Data Model** for Space and Solar Physics **interoperability**, especially within the Heliophysics Data Environment
  - Using the Data Model to create data set descriptions for all important Heliophysics data sets
  - Providing **tools and services** to assist SPASE data set description creators as well as the researchers/users
  - Working with other groups for other Heliophysics data management and services coordination as needed
- Three products:
  - SPASE Metadata Model
  - Set of Services and protocols to enable the exchange of information
  - Tools for developing and validating resource descriptions
- http://www.spase-group.org/

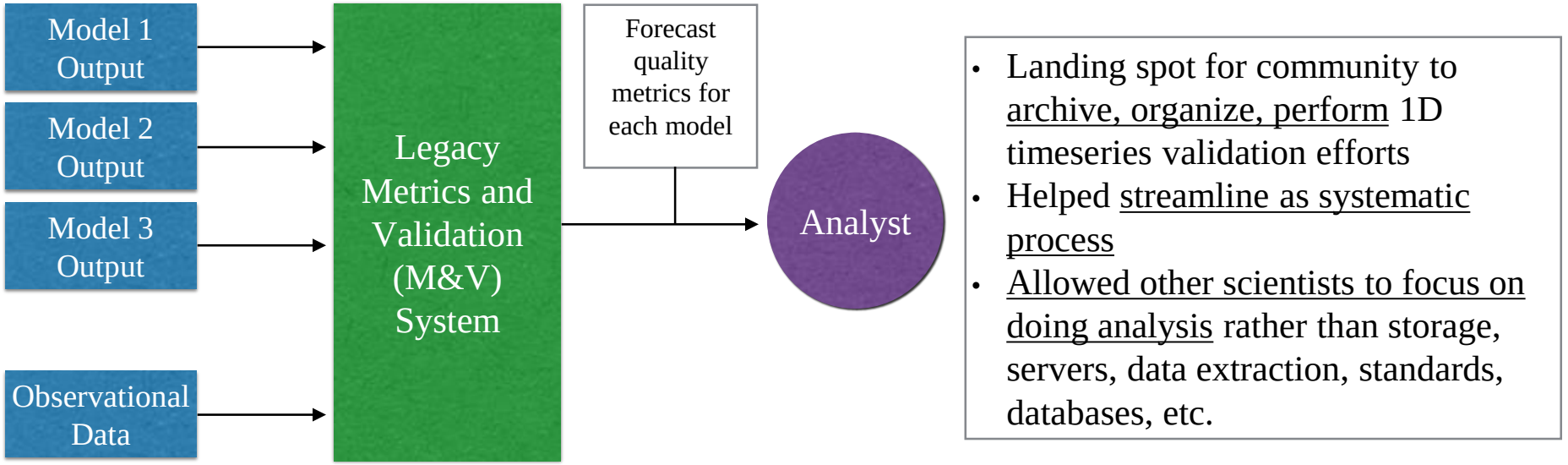
Integrated Medium for Planetary Exploration(IMPEx)

- The SPASE Simulation Extensions developed by the IMPEx project, a European Union (EU) Seventh Framework Programme sponsored project
- Describing simulations and related generated data
- http://impex-fp7.oeaw.ac.at/home.html

We need your metadata

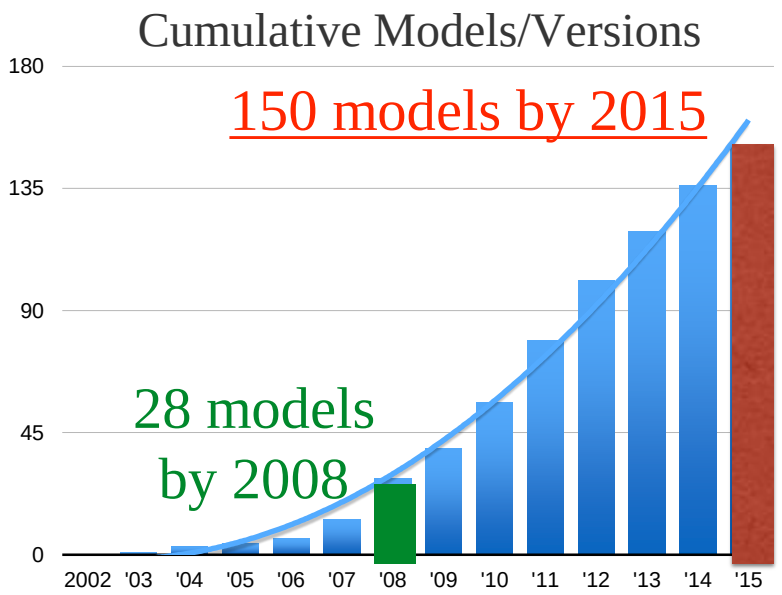


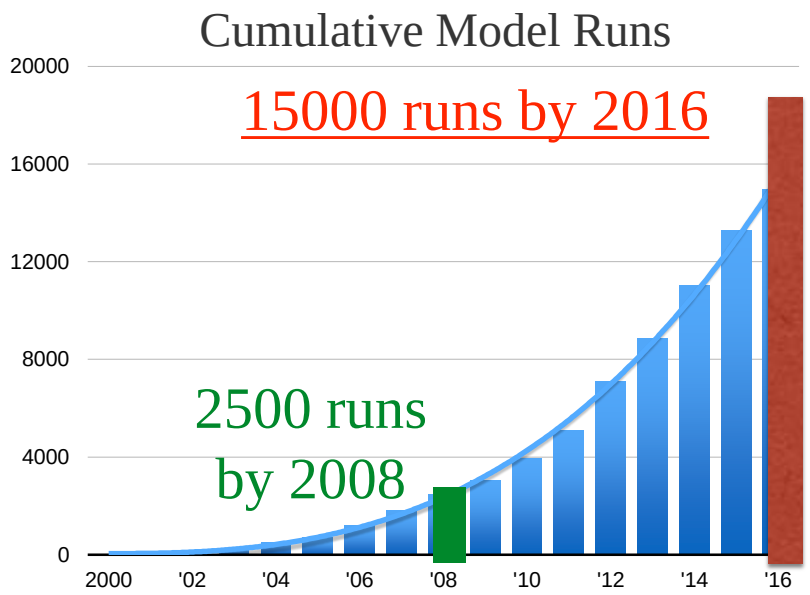
CCMC's Legacy Metrics and Validation System: Outgrown



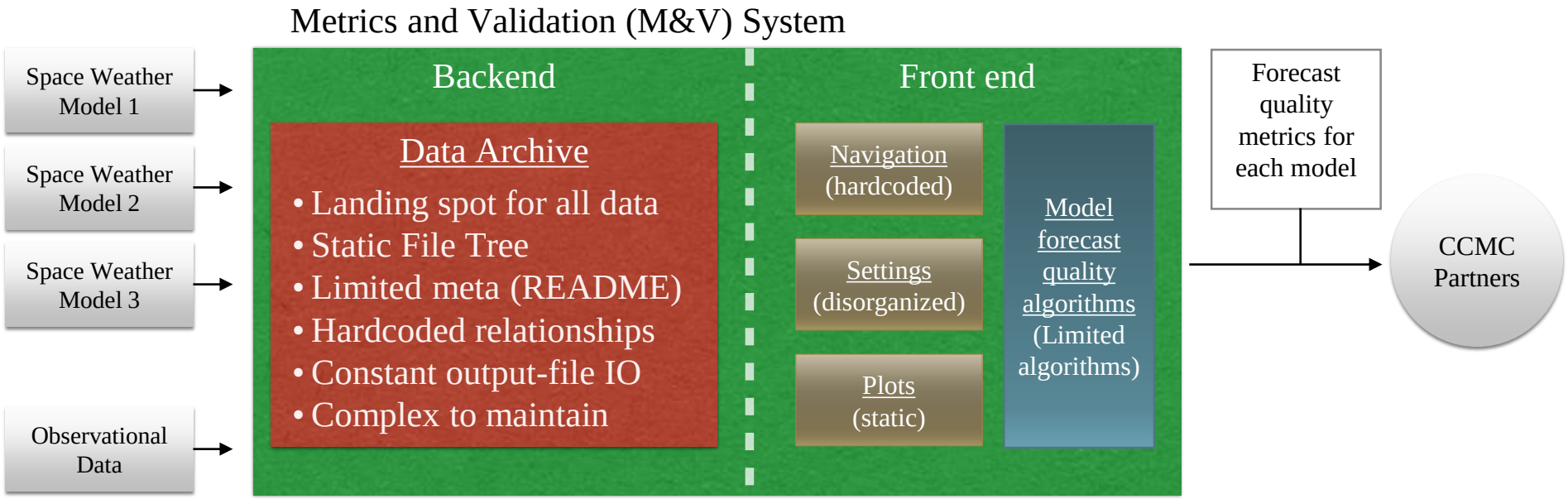
- Landing spot for community to archive, organize, perform 1D timeseries validation efforts
- Helped streamline as systematic process
- Allowed other scientists to focus on doing analysis rather than storage, servers, data extraction, standards, databases, etc.

Growth in model number and complexity





Legacy M&V components: backend scalability problem and challenge

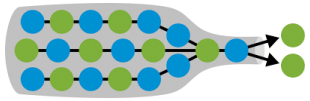


Diagnosis: (1) Data Archive was bottlenecked (2) Front end potential reengineering

Data Archive required
- Moving each model output and observational data into right file path
- Registering data in README
- Create custom parser for each format for extracting data

+

System Management
Became **tedious and manual process**



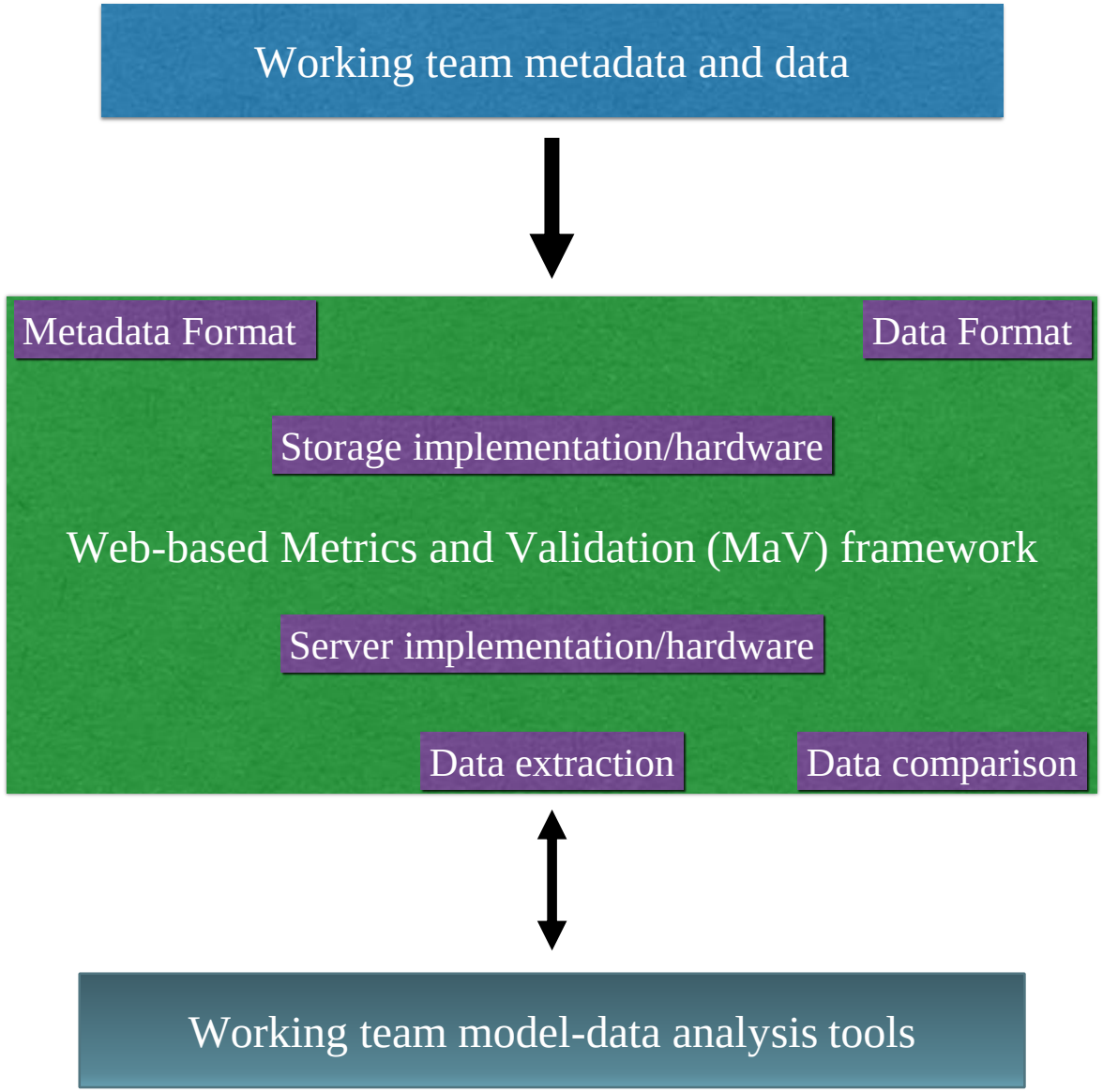
=

System Impact
- Unsustainable
- **Limited models and events available/suported** for validation
- Because hardcoded relationships, **limited types of validation analysis**
- Impeded our validation services to the community.

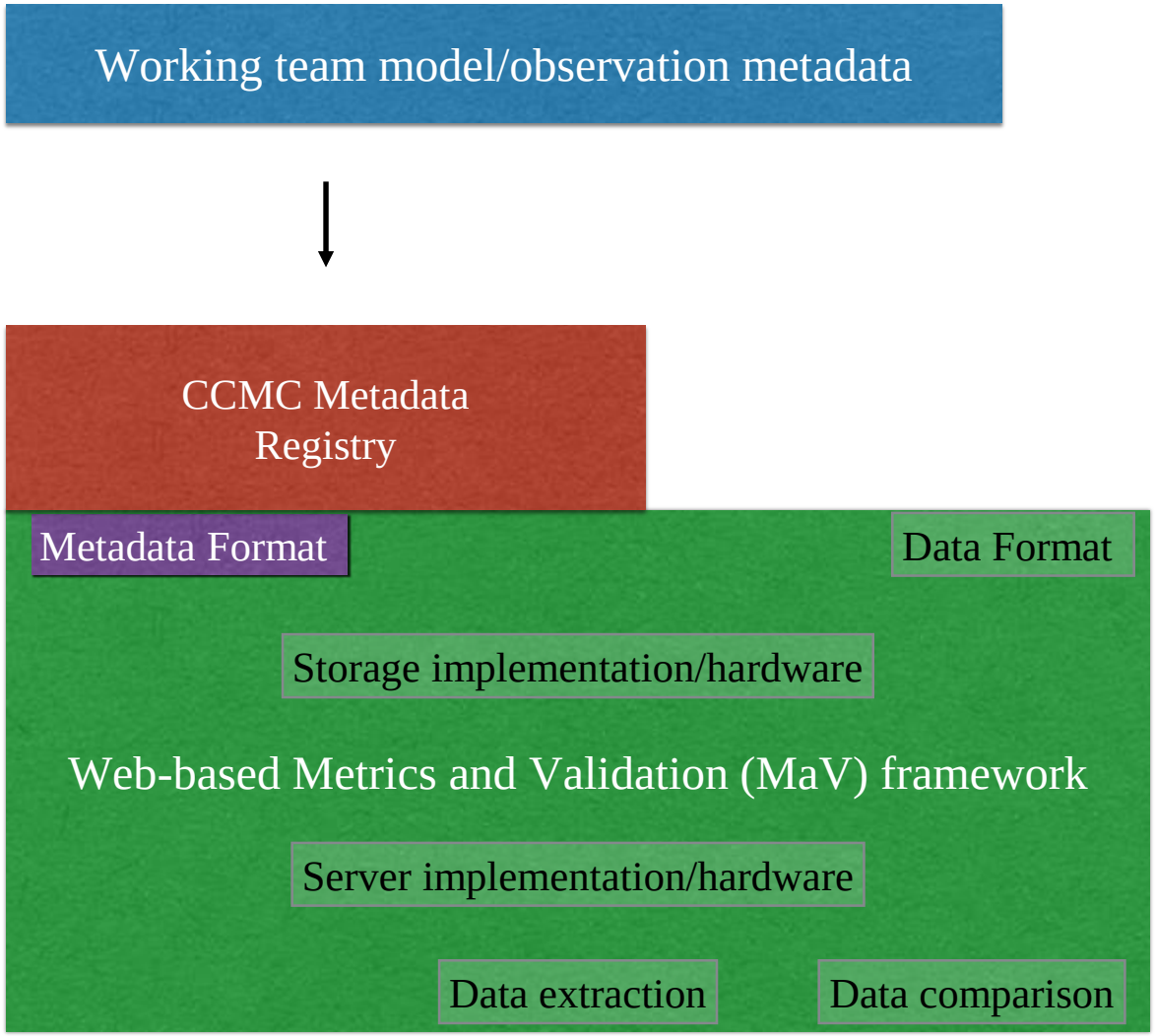
Goals of the Web-based Metrics and Validation framework

Goals

- (Just like before) Streamline validation efforts
- (How) Create a new archive that stores and provides access to all work team metadata/data
- Build on standards for optimal data discovery, utilization, and reuse
- Expose automation tools for registering new meta/data
- (Why) Reduce duplication to save working team implementation and time



Metadata format



Why? (scientists will not need to create own);

- Allows community tools all speak same language to search, discover, and utilize more data
- SPASE: Community standard, comprehensive, handle complexity of describing models/chains
- Goal: fully describe complex models/chains for data comparison, enough to reproduce output

Goals:

- Long term - scientists are registering autonomously. Power to the scientists
- Short term - work with teams closely and providing tools to help with this

Data format

Because some file formats are easier to process than others;

Assist teams in creating/choosing file formats

1D Timeseries ASII example scientists can copy:

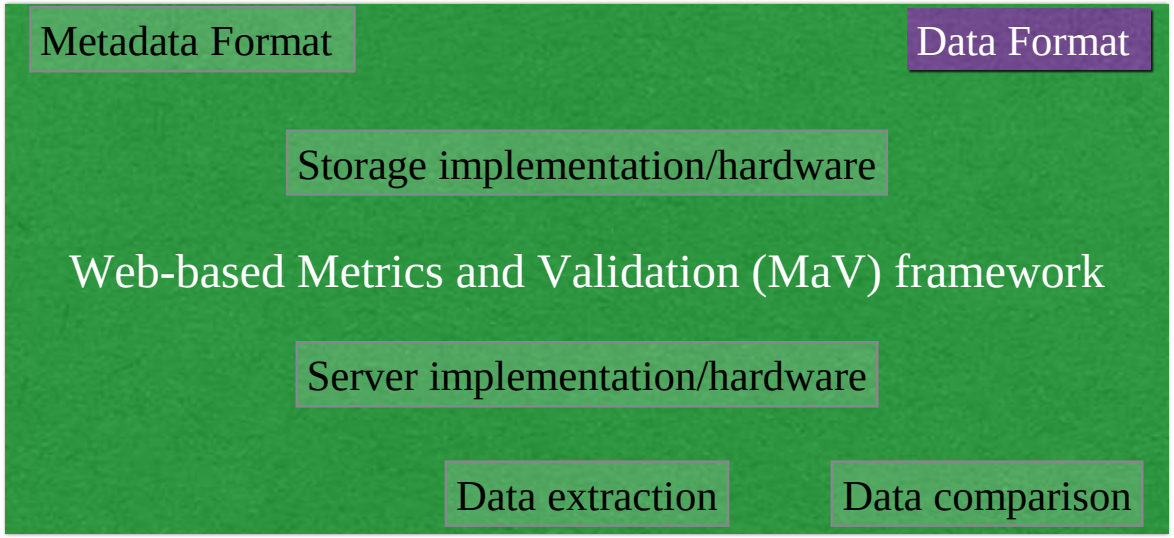
```
#Output comments
#year   mon   day   hh   mm   ss   ms    CS       foF2    foF2_median   NoDs
# [ ]    [ ]   [ ] [hr] [min] [sec] [msec] [  ]     [MHz]    [MHz]        [ ]
  2013   03    16    00   00   00   000    00100    3.850    3.590        28
  2013   03    16    00   15   00   000    00100    3.950    3.710        28
```

Configurable parser for

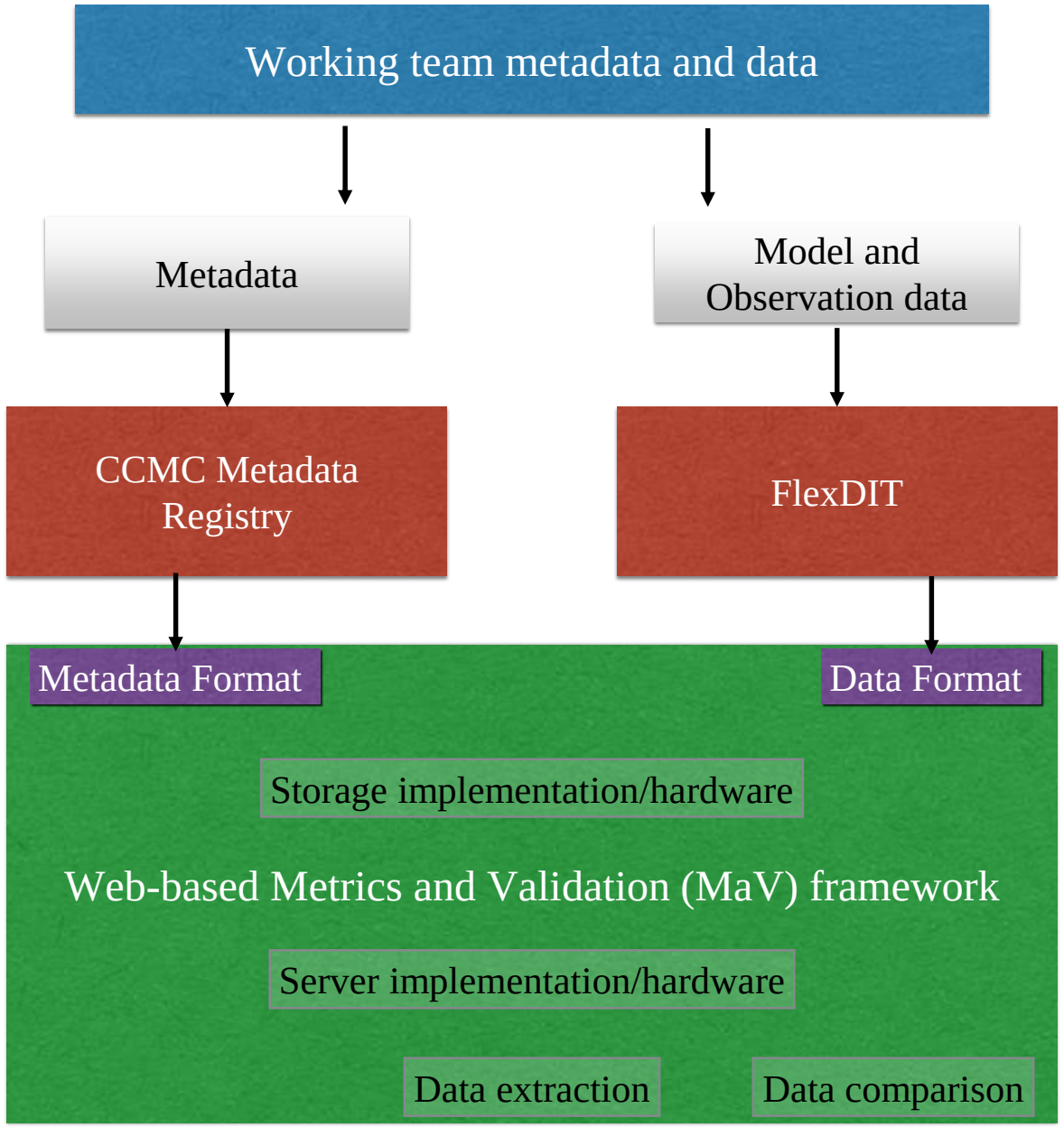
- speeding up ingestion
- handling other data formats

Flexible Data Ingestion Tool (FlexDIT)

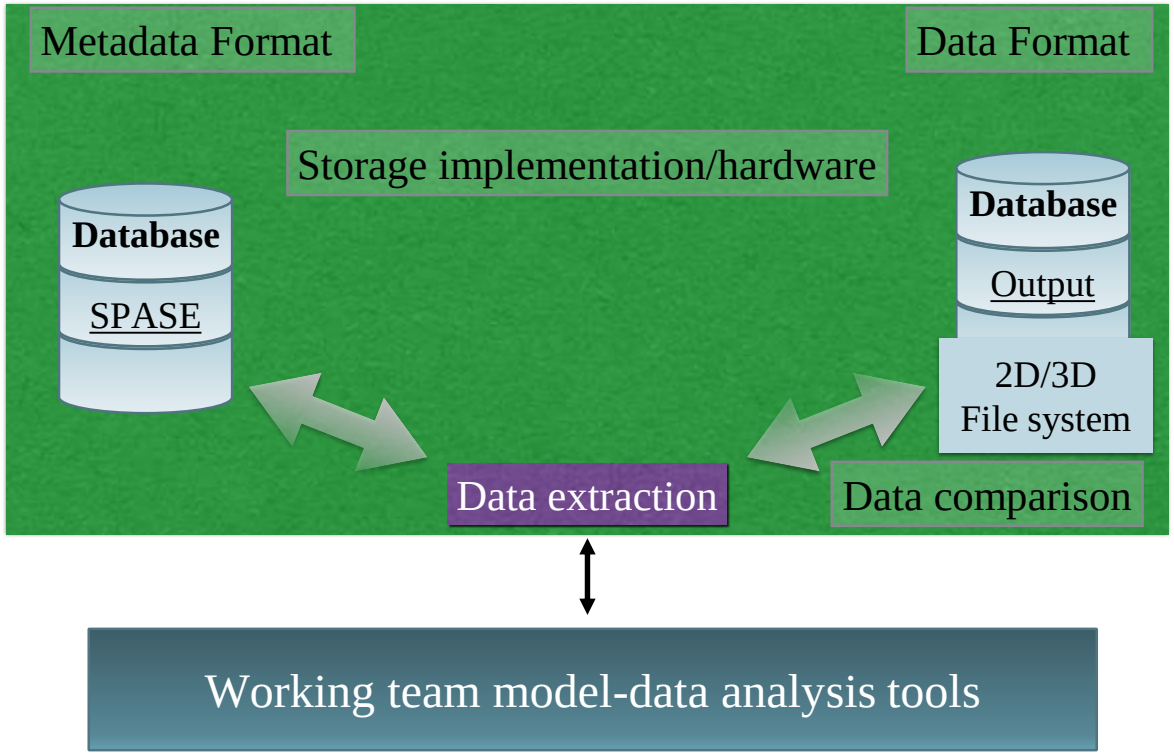
XML Descriptor File



Combine as input pipelines



Web API



Goal:

- Provide access through the Web to all working team data
- Implement an easy, standardized Web-API for teams to retrieve data for model-data comparison

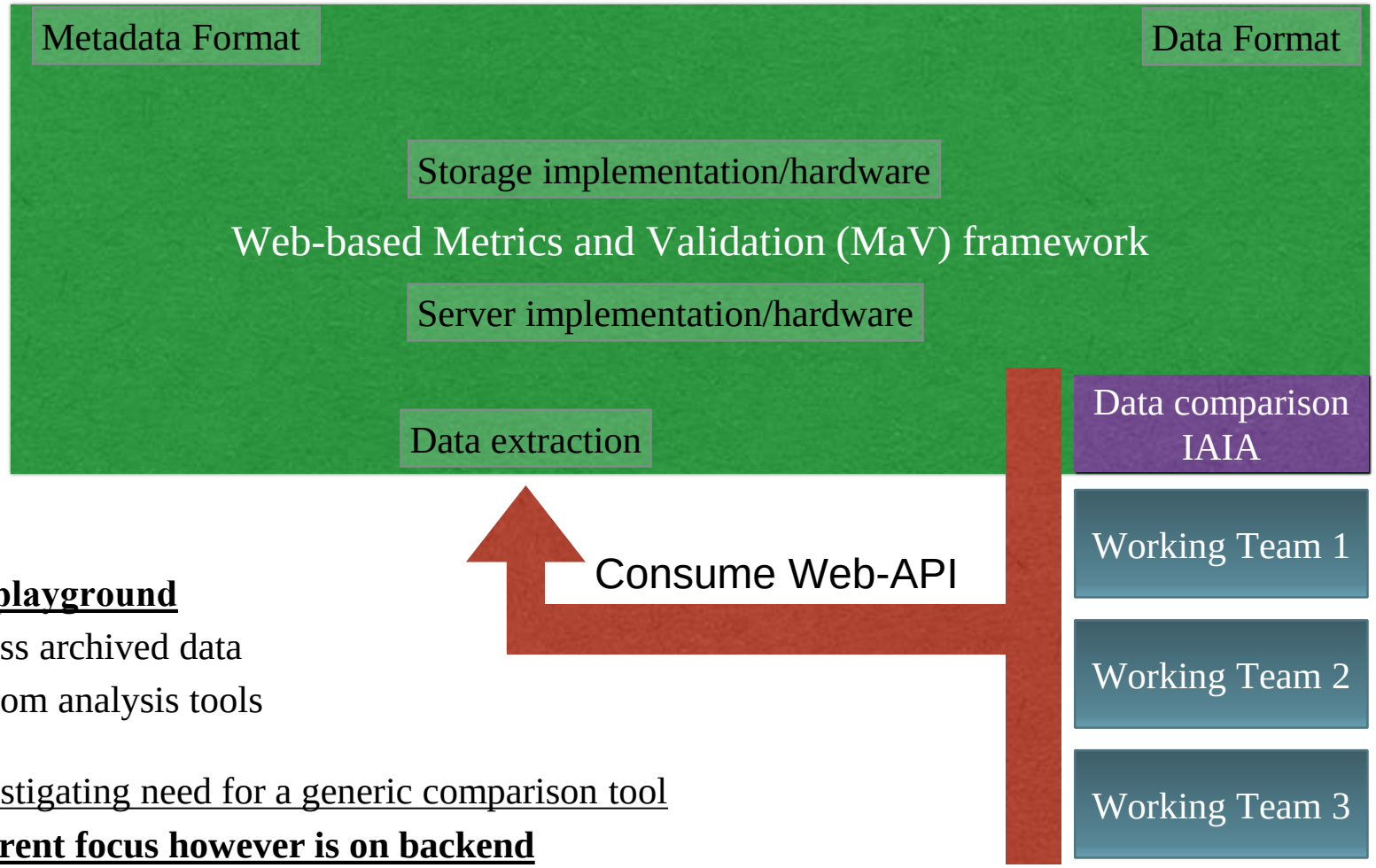
Web services to access all working team data/metadata (1D implementation initially)

```
{
  "HAPI": "1.0",
  "outputFormats": [ "csv", "binary", "json" ]
}
```

- **Parameter**: (SPASE ParameterKey)
- **Dataset**: (SPASE OutputResource ID)
- Data record: all parameters at an instance

HAPI documentation: https://github.com/hapi-server/data-specification

Model-Data comparison tools; consuming the API



Working teams' playground

- Search and access archived data
- Create own custom analysis tools

Investigating need for a generic comparison tool

Current focus however is on backend

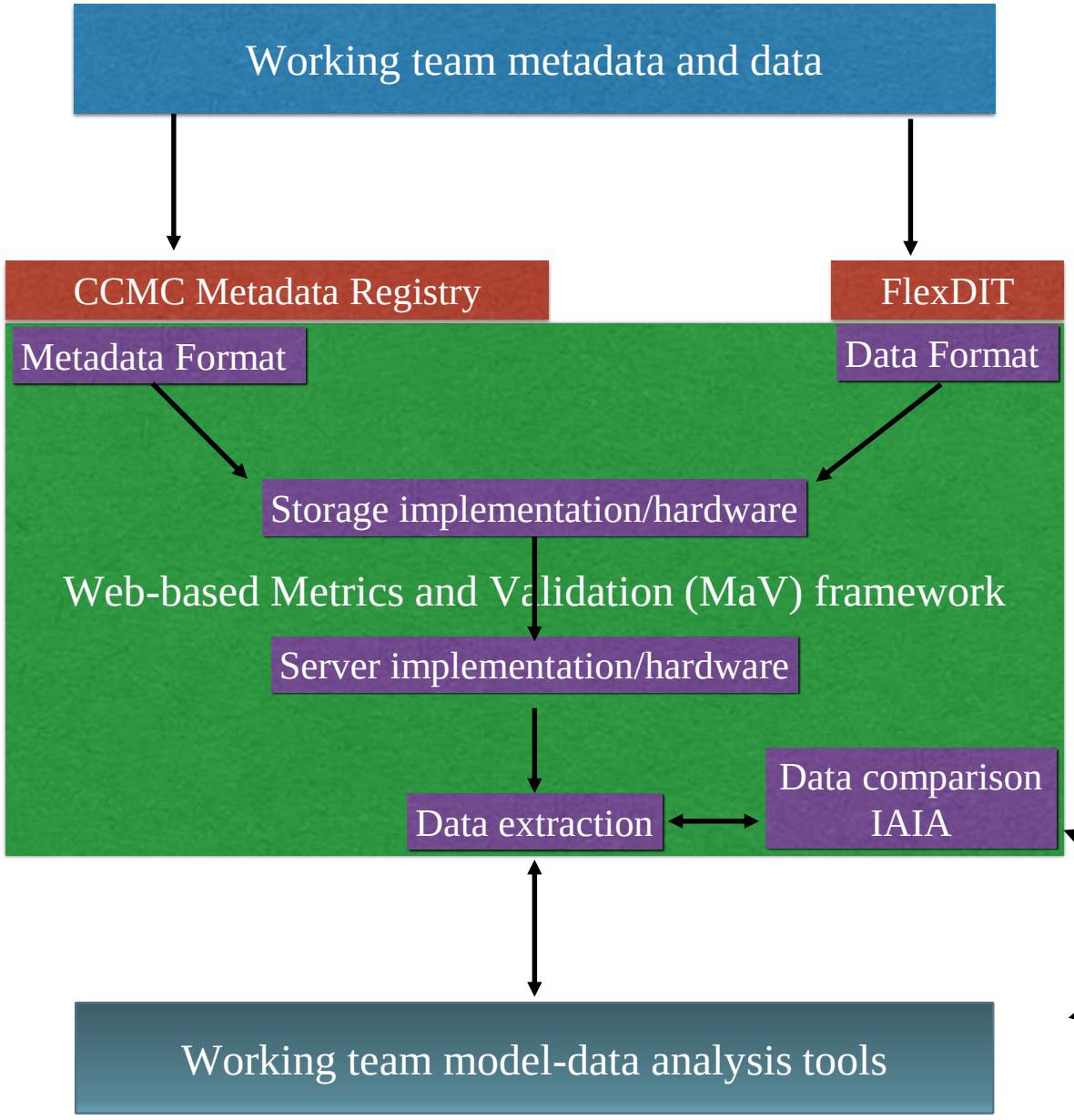
- Allow user to search catalog
- Load data sets
- Teams plug-in algorithms
- Plotting/visualize

Differences per team

- Data types (1D, 2D, 3D)
- Metrics algorithms

Frontend model-data analysis tools

Framework pipeline and benefits



- **Empower working teams** to quickly go from producing data to analyzing data
- **Reduce duplication** and **save implementation time** for teams
- Provide access via **standards for interoperability and scalability** with front-end tools
- Support **complex models and modeling chains**
- Provide **automation tools and assistance** for utilizing framework

How can we help you?

- Location: Antigua Room
- Wednesday 4:45 to 6 PM
- Thursday 4:45 to 6 PM
- What do you need from the IAIA team to support your validation effort?
- Is the current metadata model sufficient for your needs?